RUTGERS UNIVERSITY, DEPARTMENT OF COMPUTER SCIENCE

COMPUTATIONAL BIOMEDICINE IMAGING AND MODELING CENTER

VIDEO-BASED DECEPTION DETECTION USING VISUAL CUES

December 17, 2020

Madhumitha Sivaraj

Advisor: Dr. Dimitris N. Metaxas

Abstract

The act of deception is probably as old as civilization — not long after humans began communicating, they began communicating lies. Shortly after that, they probably started trying to force others to tell the truth. I built video-based deception detection models, training my task on a *Resistance Game* dataset. I developed three models (LSTM, GRU, TCN) with various aggregation techniques to accurately organize roles as deceptive and non-deceptive based on visual cues and robust facial features, such as raw pose, gaze, 1-D Facial Action Units.

1 BACKGROUND

Deception is present in our daily lives. The dictionary definition of deception is as follows: "To cause to believe what is false" [1]. Essentially, the verbal and non-verbal expressions differ from what deceivers really think and what happens. One study revealed wherever two people communicate, deception is a reality [2]. Its detection can be beneficial, not only to us individually but also to our society. One use case could extend to the public sector, where accurate deception detection could help law enforcement officers solve a crime. It can also help border control agents to detect potentially dangerous individuals during routine screening interviews.

Motivated by my interest in cognition and decision making, I wanted to explore computer vision and machine learning approaches in deception detection. This research was a contribution to an ongoing collaboration between Anastasis Stathopoulos, Ligong Han, and Dimitris Metaxas of Rutgers University, Norah Dunbar from University of California, Santa Barbara, and Judee K. Burgoon of University of Arizona. The project was funded via the US Department of Defense's Multidisciplinary University Research Initiative (MURI) grant.

My objective was to develop models and run experiments to classify different personas in the *Resistance Game* as spy or villager. The *Resistance Game* is a social role-playing game that

involves deductive reasoning. Players in the *Resistance Game* are randomly given one of two roles, spy or villager. The majority of players in a game are villagers. The spies (deceivers) know the role of everyone in the game, whereas the villagers (truth-tellers) do not. The game relies on the spies trying to hide their identity and attempting to prevent the larger group from working together to reveal the spies among the group.

2 Related Work

Previous work that inspired my project was the paper "Deception Detection in Videos using Robust Facial Features" [3]. Stathopoulos et al. propose a novel framework for non-verbal deception detection in videos that consists of two main modules: (i) a feature extractor and (ii) a video classification model.

Facial Action Unit (FAU) intensities and gaze angles are extracted from video sequences. The authors consider each one of those measurements as a 1-dimensional signal that they normalize. They concatenate those signals channel-wise and use them as input to train a predictor model. Finally, they compute the attention of the model to enable a retrospective analysis of deceptive behavior.

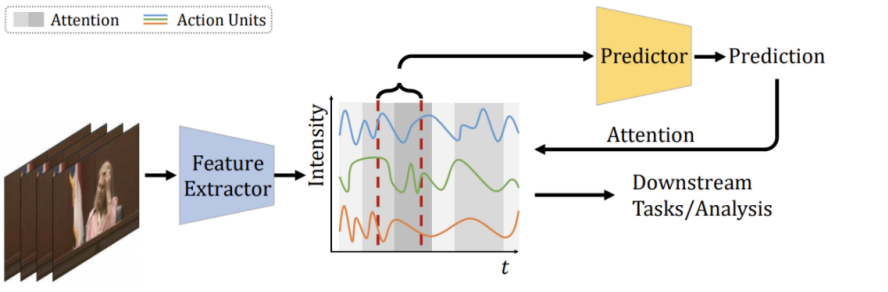

Figure 1: Illustration of feature extractor outlined in Stathopoulos et al. (2020)

The authors use a TCN for the video classification task. Because the input to the model is a waveform that carries high-level features, they could avoid using deep architectures, in contrast to other deception systems that need to extract the necessary features from the

raw video for the classification task. In the base network, the input is convolved with 128 1-dimensional kernels of size L to produce 128 1-dimensional feature maps. These feature maps are then passed through a ReLU and Batch Normalization layer. The output is processed by a dense layer implemented as 1-dimensional convolution with kernel size 1. To capture long-term dependencies in the input, the authors use residual blocks with dilated convolutions. Finally, they average-pool the feature maps and apply a fully connected layer to get the final prediction.

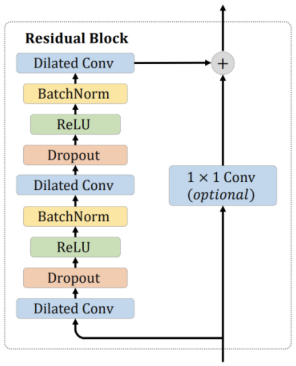

Figure 2: Illustration of residual blocks used in the proposed video classification model

The authors conclude that by using 1-dimensional FAU and gaze signals, they can train a conceptually simple, modular and powerful model that performs really well in practice. Comprehensive evaluation results illustrate that their model achieves state-of-the-art performance, even though it far less intricate than previous approaches. This highlights the usefulness of facial information for the task of non-verbal deception detection.

3 Overview

My research project builds off previous work to run more experiments, classifying different personas in the *Resistance Game* as spy or villager. I developed three models with various aggregation techniques to accurately organize roles based on visual cues.

3.1 Dataset

The critical decision points contain a set of videos involving players collected from 7 sites across the world to account for possible heterogeneity in deceptive behavior among different cultures. The data contains 17 visual cues (raw pose, gaze, FAU) extracted from videos using OpenFace. The dataset contains 3280 files, 1975 for villagers and 1305 for spies. The critical decision points involve a variety of game rounds, players, rounds, moments, and roles. However, I did not consider round and moment in my models.

Each video captures a group of 5-8 people, while playing a version of the *Resistance Game*. Players in the *Resistance Game* are randomly given one of two roles, deceivers (spies) or truth-tellers (villagers). Deceivers know who have the same role as them in the game, whereas truth-tellers are clueless. The majority of players in a game are truth-tellers; there are 2-3 deceivers. The game proceeds in rounds or missions as they are called in the game. There are 3 to 7 missions in a game. The players should nominate and elect a leader who nominates team members to go on a mission. All the players vote if that particular team should go on a mission or if a different team should be chosen. When on an mission, the team members vote in secret for the mission's success or failure. The deceivers want the mission to fail, while the truth-tellers want it to succeed. When a mission succeeds, all truth-tellers get a point, whereas when it fails, the deceivers get one point. The team with the highest score at the end of the game wins. The game relies on deceivers trying to hide their identity and attempt to prevent the larger group from working together to reveal the group's deceivers. Furthermore, it gives players many opportunities to exhibit deceptive behaviors.

The dataset used in Stathopoulos et al. (2020) involved long videos, with an average duration of 46 minutes. It contained only one label each, which made the learning procedure hard, if not impossible. For my project, we used another dataset that only contains 8 seconds, around some important timesteps. I hypothesize that the modification will yield better results. However, I fear that the current decision points are insufficient to yield high accuracy.

3.2 Approach

I built three models: LSTM (long short-term memory), GRU (gated recurrent unit), TCN (temporal convolutional networks). The models unfold to 150 timesteps. The input, x, involves 17 features (pose, gaze, FAU).

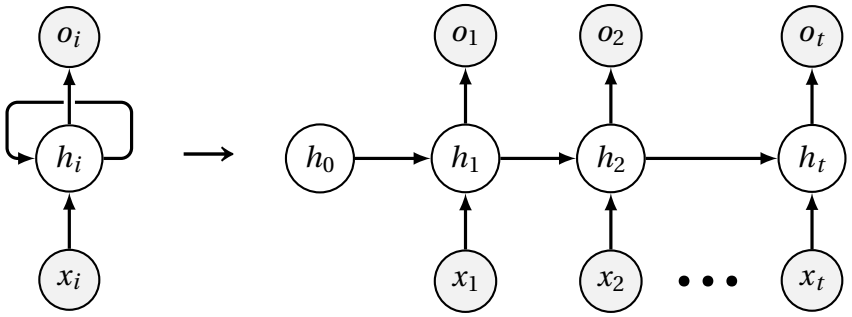

Figure 3: How the models process the input

In Figure 3, a sequence model is illustrated, where the hidden nodes are a concatenation of the previous state's output weighted by the weight matrix and the input weighted by the weight matrix. The output of the hidden state is the activation function applied to the hidden nodes. Each member of the output is produced using the same update rule applied to the previous outputs. I built the three models with aggregation techniques: (i) average-pooling, (ii) max-pooling, and (iii) last embedding encoding the whole sequence.

4 Experiments on Video-Based Deception Detection

In this section, I detail the performance of our models. I found the key frames for every video, which are the frames that contributed the most to the decision of our model. By inspecting the value of the input in those frames, we can get valuable insight into what the three models "think" that constitutes deception as a function of visual cues (FAUs intensities).

We see the results of the LSTM model with various aggregation methods on the *Resistance Game* Dataset in Table 1. The highest average classification accuracy (ACC) for the LSTM model was 58.83, with max-pooling. LSTM with max-pooling yielded the least amount of true

(TP) and false (FP) positives, and the highest amount of false (FN) and true (TN) negatives. LSTM with the last embedding being encoded as the whole sequence provided the highest result for TPs and FPs.

Table 1: Results of LSTM model on the *Resistance Game* Dataset

Aggregation	AP	TP	FP	FN	TN	ACC
Average	44.45	26.0	33.0	211.0	313.0	58.15
Max	45.67	5.0	8.0	232.0	338.0	58.83
Last	42.11	167.0	221.0	70.0	125.0	50.09

Table 2: Results of GRU model on the *Resistance Game* Dataset

Aggregation	AP	TP	FP	FN	TN	ACC
Average	42.78	194.0	276.0	43.0	70.0	45.28
Max	42.75	195.0	262.5	42.0	83.5	47.77
Last	42.94	186.5	237.5	50.5	108.5	50.60

Table 3: Results of TCN model on the *Resistance Game* Dataset

Aggregation	AP	TP	FP	FN	TN	ACC
Average	44.82	187.0	241.0	50.0	105.0	50.09
Max	44.17	191.0	243.0	46.0	103.0	50.43

Table 2 illustrates the GRU model results and shows the highest ACC for the GRU model was 50.60, with the “last embedding” aggregation technique. GRU with average-pooling yielded the most TPs and the least TNs. However, most results across all three aggregation techniques are within a smaller range than the results outlined in Table 1. The highest ACC for the TCN model was 50.43, with max-pooling, as seen in Table 3. The similarities between the results of the two aggregation techniques (average and max-pooling) are slim, with the former reporting a greater amount of TPs and FPs and the latter yielding more FNs and TNs.

Table 4 displays the comparative results on the *Resistance Game* Dataset. It reveals many interesting insights. The LSTM models with max-pooling and average-pooling yielded the

Table 4: Comparative results on the *Resistance Game* Dataset

Method	AP	TP	FP	FN	TN	ACC
LSTM (Ave)	44.45	26.0	33.0	211.0	313.0	58.15
LSTM (Max)	45.67	5.0	8.0	232.0	338.0	58.83
LSTM (Last)	42.11	167.0	221.0	70.0	125.0	50.09
GRU (Ave)	42.78	194.0	276.0	43.0	70.0	45.28
GRU (Max)	42.75	195.0	262.5	42.0	83.5	47.77
GRU (Last)	42.94	186.5	237.5	50.5	108.5	50.60
TCN (Ave)	44.82	187.0	241.0	50.0	105.0	50.09
TCN (Max)	44.17	191.0	243.0	46.0	103.0	50.43

highest ACCs. In contrast, the GRU model with max-pooling produced the lowest ACC. However, the GRU model with max and average-pooling yielded the greatest amount of TPs. Meanwhile, the LSTM model with max and average-pooling produced the least TPs.

5 Conclusion

Our results reveal that my initial belief was correct — our critical decision points is insufficient to classify with high accuracy. The next steps for this research will involve working with our collaborators at the University of California, Santa Barbara, to obtain additional critical decision points, enabling an increase in training, accuracy, and average precision. Since I do not consider the round and moment components, it may be interesting to involve these factors in future work. The model currently has 17 features, so we hope also to include additional robust, high-level features.

6 Acknowledgments

I thank my advisor Dr. Dimitris Metaxas and mentor Anastasis Stathopoulos of Computational Biomedicine Imaging and Modeling Center at Rutgers University for their unconditional support and for helping to foster my interest and growth in computer vision research.

APPENDIX

Codebase can be found at: `https://github.com/madhusivaraj/cbim_muri`.

REFERENCES

[1] Oxford English Dictionary, 1989. Oxford: Clarendon Press.

[2] Marcou, N., Zafeiriou, S., Pantic, M. (2012). A generative framsework for modeling human behavior with an application to analysis of deceptive behavior. In M. Jensen, T. Meservy, J. Burgoon, J. Nunamaker (Eds.), Proceedings of the Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium

[3] Stathopoulos A., Han L., Dunbar N., Burgoon J.K., Metaxas D. (2021) Deception Detection in Videos Using Robust Facial Features. In: Arai K., Kapoor S., Bhatia R. (eds) Proceedings of the Future Technologies Conference (FTC) 2020, Volume 3. FTC 2020. Advances in Intelligent Systems and Computing, vol 1290. Springer, Cham.